2.0 INTRODUCTION TO BOOTSTRAPPING

2.1 Introduction

There are a lot of valuable statistical methods that are practially guaranteed to work well if the data are approximately normally distributed and we are mainly concerned with linear functions of random variables. As was remarked in Chapter 1, the mean or average of a data set is a linear combination of random variables, and the central limit theorem says that we can expect means to converge on normality as the sample size increases. However, ecologists often are forced to use small samples. Very often we want to consider ratios of random variables, which are definitely nonlinear combinations, and difficult to deal with in any consistent manner. Many models of importance in ecological studies contain products, ratios and exponents, and are simply not susceptible to a standard statistical analysis in terms of the available theory.

A relatively new development in statistical methodology offers a way out of this dilemma. The technique is called "bootstrapping", which, according to Efron and Tibishirani (1993) was named from the phrase "to pull oneself up by one's bootstraps", i.e., to accomplish a physical impossiblity. Efron and Tibishirani (1993:56) note that the bootstrap was introduced by Efron in 1979, making it quite a recent development in contrast to many other statistical techniques. It was preceded by "jackknifing" which was originated by Quenouille (1956) as a way to study bias in estimators, but named by John Tukey (1958) due to its all-purpose applicability, like one's handy jackknife. A related topic is the use of the "delta method" to estimate variances for estimates based on complicated models. We will touch on these latter two methods later, but will mainly depend on bootstrapping as the principal tool for handling difficult problems.

One of the nice things about bootstrapping is that it is simple to apply, so long as one has access to a computer. Detailed application requires access to a desk computer and some knowledge of a programming language. However, bootstrapping can be done in EXCEL, as used here. There are several programming languages that can be used for bootstrapping. Most of the examples given here were also done in EXCEL, which has a random number generator in the statement RANDBETWEEN(N1,N2) where N1 and N2 represent the range of the random numbers to be generated. Be sure the ANALYSIS TOOLPAK is loaded before attempting the EXCEL versions of bootstrapping. Pull down the Tools menu and use the add-ins element to find the Analysis Toolpak. Details of use vary with the version, so you may need to use the “help” function on occasion.

2.2 The mechanics of bootstrapping

Bootstrapping is easy to apply. The process for approximating the standard error of a mean is illustrated in Fig. 2.1. An original data set containing n items (here n = 10) is randomly sampled with replacement B times, with each sample containing exactly n itemsp. Four of these B samples are shown above the original data set in Fig. 2.1. Note that an individual value from the original data set, such as 106 may appear repeatedly in a bootstrap sample. Each of the B bootstrap samples is averaged, as shown above the individual samples. This is the bootstrap _replication_. We then use these B replicate values to compute the standard error of the mean. The equation is exactly the same as that for calculating a variance, namely $s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$. However, Efron and Tibishirani use a different notation to distinguish bootstrap variables from the original data, using $x^{*1}, x^{*2}, ..., x^{*B}$ to denote the vectors containing the bootstrap samples of n observations (i.e., the four sets of bootstrap samples of 10 items each shown in Fig. 2.1). Thus the first bootstrap sample is

$$x^{*1} = (203,203,106,106,106,160,106,8,301,160).$$

The original data is represented by the vector $x = (x_1, x_2, ... ,x_n)$.

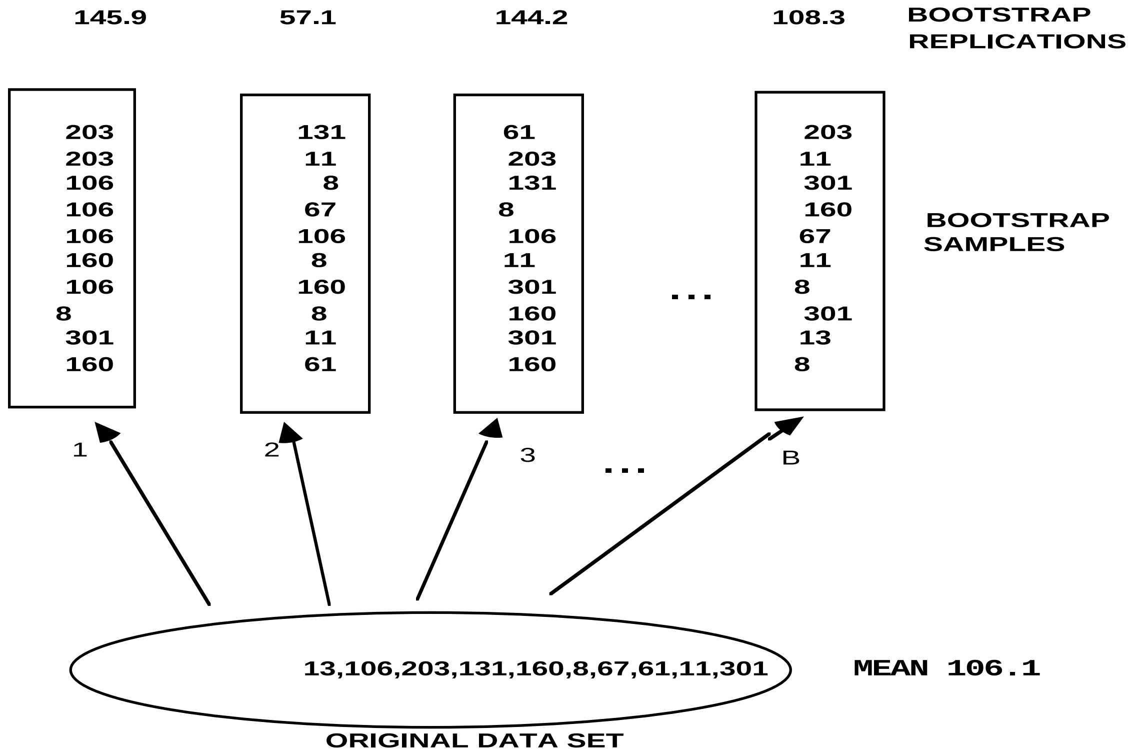

Fig. 2.1 The bootstrapping scheme for estimating a standard error. An original data set containing n items is randomly sampled B times with replacement using samples of size n. Each such bootstrap sample is averaged, and these means are used to estimate the standard error of the mean of the original data set.

The quantity $s(x^{*1})$ denotes a statistic computed from the corresponding bootstrap sample. In this case $s(x^{*1})$ is the average of the first bootstrap sample, 145.9. Using the bootstrap notation the standard error of the mean estimated by bootstrap sampling is written as:

$$\hat{se}_{boot} = \left\{ \frac{\Sigma[s(x^{*b}) - s(\cdot)]^2}{B-1} \right\}^{1/2} \qquad (2.1)$$

where the summation runs from b = 1 to B, and $s(\cdot)$ represents the mean of the bootstrap sample, i.e., $\Sigma x^{*b}/B$, where again the summation runs from b = 1 to B. The important thing to remember here is that $s(x^{*b})$ represents the mean of the b^{th} bootstrap sample, so that $s(\cdot)$ is the average of B such averages. Note, too, that the standard error of a set of random variables is computed as $s/n^{1/2}$, but here we are computing the standard deviation of a set of means and this is the standard error of the mean (i.e., don't make the mistake of dividing by the square root of B).

The first few columns of an EXCEL worksheet used to bootstrap the data of Fig 2.1 follow. The first row shows the assignment of a serial number to the original data items, while the original data appear in the second row. The next 10 rows list random numbers from 1 to 10 obtained from the statement RANDBETWEEN(1,10). The next set of numbers are random samples, with replacement, from the original data set.

ITEM NUMBER		1	2	3
DATA		13	106	203
	1	6	8	2
	2	9	7	6
RANDOM	3	1	3	10
NUMBERS	4	8	9	9
	5	2	2	1
	6	5	6	1
	7	3	8	9
	8	1	2	6
	9	7	2	2
	10	3	2	8
	1	8	61	106
BOOTSTRAP	2	11	67	8
SAMPLES	3	13	203	301
	4	61	11	11
	5	106	106	13
	6	160	8	13
	7	203	61	11
	8	13	106	8
	9	67	106	106
	10	203	106	61
SUM		845	835	638
MEAN		84.5	83.5	63.8

These are obtained by using a "table lookup" function in EXCEL. It can be explained by referring to the following formulas for the first column.

```
ITEM NUMBER       1
          DATA    13
          1       =RANDBETWEEN(1,10)
          2       =RANDBETWEEN(1,10)
RANDOM    3       =RANDBETWEEN(1,10)
NUMBERS   4       =RANDBETWEEN(1,10)
          5       =RANDBETWEEN(1,10)
          6       =RANDBETWEEN(1,10)
          7       =RANDBETWEEN(1,10)
          8       =RANDBETWEEN(1,10)
          9       =RANDBETWEEN(1,10)
          10      =RANDBETWEEN(1,10)

          1       =HLOOKUP(C3,$C$1:$L$2,2,FALSE)
BOOTSTRAP 2       =HLOOKUP(C4,$C$1:$L$2,2,FALSE)
SAMPLES   3       =HLOOKUP(C5,$C$1:$L$2,2,FALSE)
          4       =HLOOKUP(C6,$C$1:$L$2,2,FALSE)
          5       =HLOOKUP(C7,$C$1:$L$2,2,FALSE)
          6       =HLOOKUP(C8,$C$1:$L$2,2,FALSE)
          7       =HLOOKUP(C9,$C$1:$L$2,2,FALSE)
          8       =HLOOKUP(C10,$C$1:$L$2,2,FALSE)
          9       =HLOOKUP(C11,$C$1:$L$2,2,FALSE)
          10      =HLOOKUP(C12,$C$1:$L$2,2,FALSE)

SUM               =SUM(C14:C23)
MEAN              =C25/10
```

The statement HLOOKUP(C3,$C$1:$L$,2,FALSE) specifies a horizontal lookup table (VLOOKUP permits a vertical lookup table). The first entry is the column entry for the value to be looked up in the table, i.e., C3 denotes a random number entry, for which we need to find the corresponding entry in the original data row. The lookup table is specified by the array, $C$1:$L$1 in which the first row is the index value corresponding to a data entry in the next row. The subsequent value in HLOOKUP is the row containing the data to be returned by the HLOOKUP function, and the final entry ("FALSE") insures that the function returns the exact value required (using "TRUE" would permit returning the value nearest in numerical magnitude to a lookup entry). As with any of the more complex functions in EXCEL, a little practice will make the role of the individual entries clear. An important proviso with the HLOOKUP and VLOOKUP functions is that the lookup table must be in the first rows (or first columns for VLOOKUP) of the spreadsheet. The last entries above give the sums and means of the bootstrap samples. The means are used in eq.(2.1) to calculate the bootstrap standard error. Readers should understand that the example used here is mainly intended to demonstrate the approach. The best estimate of a standard error of a set of numbers is that calculated by the usual formula, i.e., from

$$\mathrm{S.E.}^2 = \frac{\Sigma(x_i - \bar{x})^2}{n(n-1)} \ .$$

Bootstrapping is used to calculate standard errors for more complex functions, for which a direct estimate of a variance is not available from statistical theory.

Example 2.1. The original data of Fig. 2.1: 13,106,203,131,160,8,67,61,11,301 represent data on survival times in days. They can be considered to come from an experiment on the effect of some treatment on survival of experimental animals (whereupon there should be a corresponding set of data from a control group) or the survival times of a set of radio-tagged wild animals. It is a small sample, but this is common, inasmuch as there is increasing public pressure to reduce experimental use of live animals, and collecting data from wild animals is expensive and can be quite difficult. We would thus like to extract as much information as possible from the data. The data in Table 2.1 are from an EXCEL worksheet that computes the bootstrap standard error. It shows the first 10 columns of a total of 50, which is likely the minimum size that should be used to demonstrate behavior of bootstrapping. In preparing such spreadsheets, one should change the calculation mode from automatic to manual (in the TOOLS menu, under OPTIONS or PREFERENCES depending on the version of EXCEL) while building the worksheet. The calculate command can then be used to see how the result varies from run to run.

2.3 Empirical probability distributions

The probability distribution of a random variable, X, is any complete description of the probabilistic behavior of x. In coin-tossing with a "fair" coin, there are two possibilities, each occurring with probability 1/2. In rolling a die, there are 6 outcomes, each having $\Pr\{x = k\} = 1/6$ for $k = 1,2,3,4,5$, or 6. It is convenient to define the sample space, S_X, as a list of possible outcomes. Thus for a fair die, $S_X = \{1,2,3,4,5,6\}$ and we assign probability 1/6 to each event in the sample space. Consider the binomial distribution which assigns a probability to each sample point in the sample space {0,1,2,3, ... , k, ... , n} but these probabilities depend on the parameter, p, of the distribution. The binomial distribution is:

$$\text{Prob}\{x_i = k\} = f_k = \binom{n}{k} p^k(1-p)^{n-k} \tag{2.2}$$

where $\binom{n}{k}$ is evaluated as $\frac{n!}{(n-k)!k!}$, in which, for example, 5! (read as "five factorial") is calculated as 5x4x3x2x1 = 120.

Table 2.1 Sample from bootstrapping the data of Fig. 2.1.

ITEM NUMBER		1	2	3	4	5	6	7	8	9	10
DATA		13	106	203	131	160	8	67	61	11	301
	1	7	8	1	6	5	4	6	2	8	6
	2	10	10	3	8	10	10	2	4	3	2
RANDOM	3	2	7	4	10	7	4	3	5	10	10
NUMBERS	4	4	1	4	4	5	6	8	10	3	7
	5	8	6	1	6	7	8	5	3	1	4
	6	9	7	9	4	4	9	1	7	6	6
	7	1	5	6	9	10	3	5	3	9	7
	8	1	8	3	2	10	10	10	1	1	7
	9	5	2	6	6	9	10	1	10	10	6
	10	2	6	3	8	7	3	2	2	2	6
	1	67	61	13	8	160	131	8	106	61	8
BOOTSTRAP	2	301	301	203	61	301	301	106	131	203	106
SAMPLES	3	106	67	131	301	67	131	203	160	301	301
	4	131	13	131	131	160	8	61	301	203	67
	5	61	8	13	8	67	61	160	203	13	131
	6	11	67	11	131	131	11	13	67	8	8
	7	13	160	8	11	301	203	160	203	11	67
	8	13	61	203	106	301	301	301	13	13	67
	9	160	106	8	8	11	301	13	301	301	8
	10	106	8	203	61	67	203	106	106	106	8
SUM		969	852	924	826	1566	1651	1131	1591	1220	771
MEAN		96.9	85.2	92.4	82.6	156.6	165.1	113.1	159.1	122	77.1

For convenience in discussing bootstrapping, we can describe a probability distribution as $F\{f_1,f_2,f_3, \ldots ,f_k \ldots ,f_N\}$ where f_i is the limiting frequency of the i^{th} event. For a single die, we infer that f_i = 1/6, and would expect to eventually come very close to that value, given enough rolls of the die. If we determine f_i from observations, then it can be considered to be an empirical probability distribution. Instead of rolling dice, we can set up a spreadsheet using RANDBETWEEN(1,6), copy this down through, say, 1,000 cells, and tabulate the outcomes by using the histogram function in the data analysis menu under TOOLS. This gave the following results:

Bin	*Frequency*	*Proportion*	*F(x)*
1	181	0.1810	0.1810
2	179	0.1790	0.3600
3	162	0.1620	0.5220
4	172	0.1720	0.6940
5	152	0.1520	0.8460
6	154	0.1540	1.0000
TOTAL	1000		

The column under proportion gives the empirical frequency distribution, with the corresponding cumulative frequency distribution being shown under $F(x)$.

It is necessary to note that, in mathematical statistics, $F(x)$ represents the cumulative probability distribution function, $F(x_o) = Pr\{x <\_x_o\}$.The table above provides an example. We will use the description $\hat{F}\ \{f_1,f_2,f_3, ... ,f_k ... ,f_N\}$ for a finite number of events as a handy way to represent an empirical probability distribution in discussing bootstrapping. The "hat",^, over a symbol means that the quantity is an estimate of the true, but unknown, value, F. The cumulative, $F(x_o) = Pr\{x \leq x_o\}$, will mainly be used here in simulations. It is important to remember that the sum of the frequencies in $F\{f_1,f_2,f_3, ... ,f_k ... ,f_N\}$ is always unity, described as $Pr\{x \varepsilon S\} = \Sigma\ f_k = 1$, where "$\varepsilon$" means "contained in".

2.4 Sample sizes for bootstrapping

Bootstrapping is a resampling procedure, that is, we take repeated samples of the original data set, calculate values of some statistic $s(x^*)$ and use these to infer something about the true, but unknown value of some parameter. In the example used thus far, the statistic was the standard error of the mean. How many bootstrap replications are needed? Efron and Tibishirani [1993: Eq.(6.9)] give a formula for examining the effect of varying sample size, but also indicate that, in their experience, $B = 200$ is usually adequate for estimating the standard error, while $B = 50$ may provide useful information. In the problems I have dealt with thus far by using bootstrapping, I tend to use $B = 100$ for exploring data and debugging programs, and $B = 1{,}000$ or $2{,}000$ for the published result. With desktop computing so cheap, one might as well resort to "overkill" unless the statistic being bootstrapped is very complicated and requires a lot of computing time. However, this choice of $B \geq 1{,}000$ is also largely driven by the fact that larger samples are needed to compute confidence limits by bootstrapping, as we'll see in the next section. When making calculations using EXCEL in some older versions one can only get about 250 bootstraps in the horizontal plane, so it is desirable to use VLOOKUP and set up the table in the vertical plane, whereupon it is possible to get 2,000 or more bootstraps for confidence limits. If more bootstraps are needed, one can copy off data to another sheet and recalculate, copy off those results, and recalculate again. Large samples can thus be obtained. However, as noted above, 2,000 is usually adequate for confidence limits.

2.5 Percentile confidence limits

Calculating confidence limits by bootstrapping can be extremely simple, if the percentile method is used. Follow the same process demonstrated in Fig. 2.1, generating at least 1,000 bootstraps (I tend to use 2,000 if computing doesn't take too long), store the data in a file, arrange it in numerical order, and count in $\alpha B/2$ observations from both ends, where α is the chosen "significance level". These are the percentile confidence limits. Although there is nothing in the underlying theory that dictates a choice, most biologists tend to use $\alpha = 0.05$, for 95% confidence limits.

To accomplish this in EXCEL, use VLOOKUP and put the data in the first 2 columns, the same number of random numbers in the third row as there are data points, and the corresponding lookup values in the fourth row. Calculate the function being bootstrapped in subsequent rows. You can then order the data and count in from both ends for confidence limits. In this case, the function being bootstrapped is simply the mean. Confidence limits are obtained by ordering this column (use SORT in the DATA menu of EXCEL) and counting in from the ends of the ordered data. To use α = 0.05 on 1,000 bootstraps, one would count in 25 observations from each end. In this example, the approximate 95% confidence limits were 55.4 to 169 around the mean of 106.1 of the original data, based on 1,000 bootstraps. Using a BASIC program to do the bootstrapping is faster and requires less effort once the programming is done. Results of 2,000 bootstraps from a BASIC program (Fig. 2.2) gave confidence limits of about 55-164.

As noted previously, EXCEL will accomodate at least 2,000 bootstraps in the vertical arrangement. However, if an older (and thus slower) computer is used, it may be best to do only 200 bootstraps at a time. That is, set up the operation as shown on the attached sheet, run 200 bootstraps, and copy the results to a second worksheet. Do this 5 times (or 10 if you want 2,000 bootstraps) and then order the data on the second worksheet to locate the confidence limits.

Students should review normal theory confidence limits in the statistics text of their choice at this point. Under normal theory, we would calculate a standard error of the mean of the original data of Fig. 2.1 (mean = 106.1), getting s = 95.33, and S.E. = $95.33/(10^{1/2})$ = 30.14, and calculate 95% confidence limits of $\underline{+}$ 1.96 SE. I tend to use 2 rather than 1.96 for convenience, and a little extra margin. Using 2 S.E. gives approximate 95% limits of 46 to 166. Survival data generally follow a highly skewed distribution, and the sample variance tends to vary appreciably. In this case, the limits are so wide that the data don't give us a very good notion of average survival time.

Statistics books recommend transforming skewed data in order to approximate normality. One then produces normal-theory confidence limits as above, and transforms back to the original scale. It can, however, be a considerable chore to find a normalizing transformation suitable for the data at hand. Further, the small sample of Example 2.1 simply does not supply enough information to evaluate possible transformations. It is thus reassuring that Efron and Tibishirani (1993: Chap. 13) indicate that the percentile method automatically supplies limits that would be obtained under normal theory if we knew the proper transformation to normalize the data.

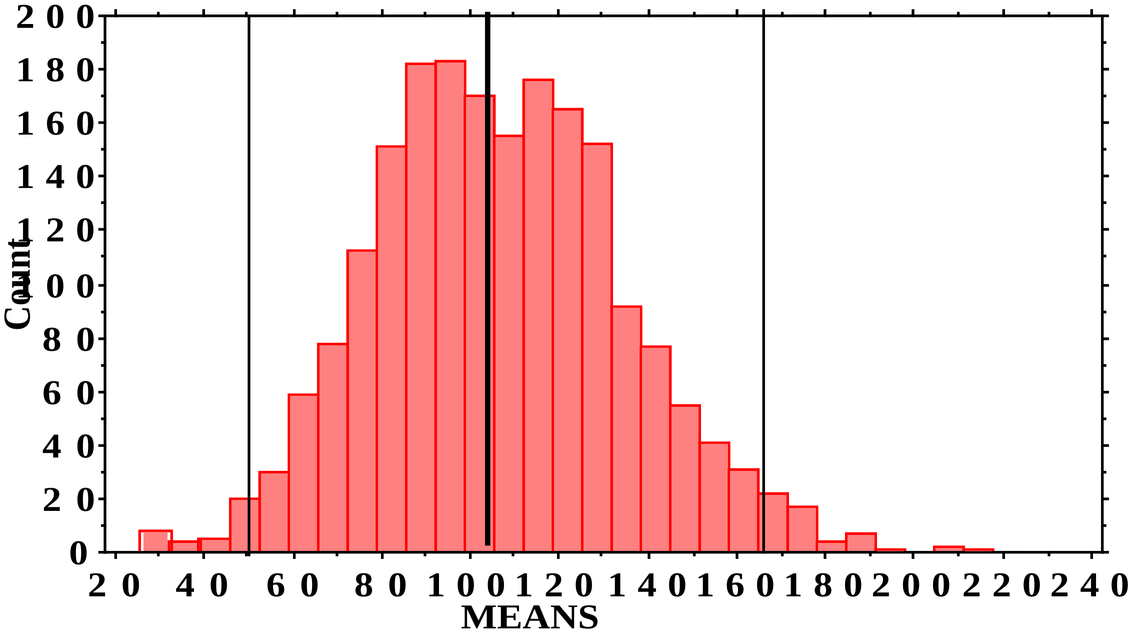

Fig. 2.2 Frequency plot of 2,000 bootstraps of the original data of Fig. 2.1, showing mean of the original data (heavy central vertical line), and 95% confidence limits (lighter lines to right and left) from the bootstrap percentile method for calculating confidence limits. These limits are at about 55 and 164 for the data shown. Note that the appearance of these graphs will vary somewhat as this is a sampling procedure.

2.6 Regression models and parametric bootstrapping

Regression models provide extremely valuable tools in ecological studies. Many investigators use regressions without giving much thought to the matter, and may thus report some erroneous results without realizing that this is possible. Regression models are classified as linear and nonlinear. Linear models are most commonly used, with the main example being $y = \alpha + \beta x$, where α denotes the "intercept" and β the "slope". Ecologists also use multiple regression models with two or more x-values, e.g., $y = \alpha + \beta_1 x_1 + \beta_2 x_2$, and may also use multiple regression models like $y = \alpha + \beta_1 x_1 + \beta_2 x_2^2$. These are both linear models, being "linear in the coefficients", but a version like $y = \alpha + \beta_1 x_1 + \beta_2 x_2^{\gamma}$ is nonlinear. A frequently encountered nonlinear model is $y = \alpha e^{-\beta x}$. This model can, however, be transformed into a linear model by taking logarithms (usually to base e) giving $\log_e y = \log\alpha - \beta x$. The model $y = \alpha + \beta_1 x_1 + \beta_2 x_2^{\gamma}$ is said to be intrinsically nonlinear, inasmuch as a simple transformation will not convert it to a linear version (unless, of course, we know or assume we know γ). Dealing with intrinsically nonlinear models can be difficult, and they are most often fitted with nonlinear least-squares. Programs are available for fitting by nonlinear least-squares.

Regression models may be bootstrapped in exactly the same way as shown in Fig. 2.1, except that now the original data will consist of x,y pairs, and the statistic computed from bootstrap replications consists of paired estimates of α and β, rather than the mean as used in the example of Fig. 2.1.

How these sets of paired estimates are treated depends on the purpose of the study. Often the main interest is in estimates of β, but we may also want to set confidence limits on an estimate of some value of x computed from the estimates of α and β. Texts on regression analysis are available; one of the more widely used is that of Draper and Smith (1998), and most basic statistics texts give a good deal of attention to regression models. To set confidence limits on some regression estimate by bootstrapping, one simply needs to follow the procedure presented above, with the "statistic" being the estimate of interest in the study at hand.

The chief problem for ecologists in this approach is the usual one -- small sample sizes. With smallish samples, bootstrapping pairs may give some strange and variable results. We will thus need to consider parametric bootstrapping. The procedure again is simple. One fits a regression model to the original data, calculates residuals about the fitted line, and bootstraps the residuals. Consider the result of fitting a simple linear regression to n original pairs of x,y observations. The outcome is a fitted regression line, denoted as $\hat{y}_i = a + bx_i$, where a and b represent the estimates of α and β, and there are n pairs of original data. The residuals about regression are calculated as:

$$e_i = \hat{y}_i - y_i \qquad (i = 1,2,3, ..., n) \qquad (2.2)$$

where $\hat{y}_i$ is calculated from the fitted line, $\hat{y}_i = a + bx_i$. We now bootstrap the residuals, taking repeated random samples with replacement of n observations from the residuals, add these residuals to the fitted regression line to get a new set of n values of y_i. Combined with the original set of x-values (unchanged throughout) these new pairs constitute the bootstrap samples of Fig. 2.1. We then calculate the bootstrap replication by fitting a new regression line to the bootstrap sample. Of course, if we are only interested in, say, the slopes, b, then only that calculation needs to be carried out. The only tricky part is to remember that the new values of y_i are computed from the i^{th} value of x_i, so that the same residual (e_i) may be associated with several values of x_i, depending on the random selection. That is, the new set of y_i values is computed from:

$$y_i = a + bx_i + e_i \qquad (i = 1,2,3, ..., n) \qquad (2.3)$$

with a and b coming from the regression line fitted to the original data and the values of e_i come from a random sample with replacement of the n data points generated by eq.(2.2). Students should repeat the calculations shown below to fix the scheme in mind. EXCEL will produce fitted values (check the “residuals” box) which can be used to construct bootstrap samples by adding them to the e_i.

Example 2.2 Parametric regression bootstrapping.

For simplicity, we will suppose that we want 95% confidence limits on the slope, β, of a regression line. The slope estimate is calculated as:

$$b = \hat{\beta} = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2} \qquad (i = 1,2,3, ..., n) \qquad (2.4)$$

and the intercept estimate is $a = \hat{a} = \overline{y} - b\overline{x}$.

STEP 1 Compute a regression line from the original data (EXCEL does this for you):

$$b = = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\Sigma(x_i - \overline{x})^2} = \frac{325.181239}{630.5} = 0.5157$$

$$a = \overline{y} - b\overline{x} = 16.725 - (0.5157)(20.5) = 6.1525$$

$\hat{y} = 6.1525 + 0.5157x$ Regression line from original data

STEP 2 Calculate the deviations $e_i = \hat{y}_i - y_i$

Original data

i	x_i	y_i	$\hat{y}_i$	$e_i = \hat{y}_i - y_i$
1	10	12.672	11.31	1.362
2	12	8.9391	12.3415	-3.402
3	14	13.934	13.373	0.561
4	15	16.377	13.8888	2.488
5	17	13.252	14.9203	-1.668
6	21	19.121	16.9833	2.137
7	23	17.821	18.0148	-0.194
8	28	18.879	20.5936	-1.715
9	30	21.047	21.6251	-0.578
10	35	25.213	24.2038	1.009

STEP 3 Draw random samples of 10 with replacement from the e_i:

i	e_i	Random samples with replacement from the e_i					
1	1.362	-0.1936	0.5607	-1.6681	-0.5776	-3.4024	1.0091
2	-3.402	-0.5776	2.4879	-0.1936	-0.1936	1.3618	-3.4024
3	0.561	-1.7150	-1.7150	1.0091	-3.4024	1.0091	0.5607
4	2.488	0.5607	1.3618	2.4879	-0.1936	-3.4024	1.3618
5	-1.668	1.3618	-0.1936	-1.7150	-0.1936	1.0091	2.1373
6	2.137	-1.6681	-1.6681	1.3618	-0.5776	1.0091	1.3618
7	-0.194	2.1373	-0.1936	0.5607	-1.6681	-3.4024	-3.4024
8	-1.715	-0.1936	1.0091	2.4879	-1.7150	-3.4024	-3.4024
9	-0.578	-1.7150	-0.5776	1.3618	1.3618	2.4879	0.5607
10	1.009	1.3618	-1.6681	1.3618	1.0091	-1.7150	2.1373

STEP 4 Add the random samples of e_i to the predicted regression line to obtain new sets of y_i:

i	x_i	$\hat{y}_i$	$\hat{y}_i$ + random samples with replacement from the e_i					
1	10	11.31	11.1164	11.8708	9.6419	10.7325	7.9076	12.3192
2	12	12.342	11.7640	14.8294	12.1479	12.1479	13.7033	8.9391
3	14	13.373	11.6580	11.6580	14.3822	9.9706	14.3822	13.9338
4	15	13.889	14.4495	15.2506	16.3766	13.6952	10.4864	15.2506
5	17	14.92	16.2821	14.7267	13.2053	14.7267	15.9294	17.0576
6	21	16.983	15.3152	15.3152	18.3451	16.4057	17.9924	18.3451
7	23	18.015	20.1521	17.8212	18.5755	16.3467	14.6124	14.6124
8	28	20.594	20.3999	21.6027	23.0814	18.8786	17.1911	17.1911
9	30	21.625	19.9101	21.0475	22.9868	22.9868	24.1129	22.1858
10	35	24.204	25.5656	22.5357	25.5656	25.2130	22.4888	26.3411

STEP 5 Calculate regression slopes for each of these new sets of "data", using the same set of x_i. This gives the values, 0.542,0.443,0.608,0.582,0.513, and 0.520. In practice, of course, one would calculate a large set of such estimates, 1,000 or more. The frequency distribution of these values then provides the basis for confidence limits, as calculated previously for means.

A generalized summary of the steps in parametric bootstrapping is as follows:

1) Compute estimates of the parameters of the model from the original data. In this case, the regression coefficients, a and b.

2) Calculate deviations, $e_i = \hat{y}_i - y_i$, between the observed data (y_i) and the fitted model ($\hat{y}_i$).

3) Draw B (at least 1,000 for confidence limits) random samples of n with replacement from this set of deviations.

4) Add these deviations to the $\hat{y}_i$ to create the bootstrap replications.

5) Compute parameter estimates from each of these B sets of data.

6) Arrange these B estimates in a frequency distribution and count in αB/2 observations from each end to obtain (1-α)% confidence limits.

Calculations can be carried out in EXCEL by using the same arrangement as used in Sec. 2.5 to get confidence limits on a mean. The data to be bootstrapped are now the deviations from regression, and the bootstrap operation proceeds in exactly the same manner. However, another stage has to be incorporated in which the bootstrapped deviations are added to the predicted regression. These new regression values are then used to estimate the parameters of the regression equation. In the present example, only the slope is calculated. This can be done by using the SLOPE function, which returns the slope of two arrays. The x-values are the original values, while the y-values are those in the body of the table. The 1,000 slope values were then ordered, and approximate 95% confidence limits obtained by counting up and down 25 entries. The limits obtained from the EXCEL calculation (B = 1,000) were 0.375 and 0.648. A calculation using a program written in BASIC were 0.377 to 0.652. A plot of the results of 2,000 bootstraps computed by the BASIC program appears in Fig.2.3.

Students should review normal theory regression calculations in Chap. 1.0 or in a statistics textbook. The variance about regression is calculated as follows:

$$s^2 = \frac{\Sigma(y_i - \hat{y}_i)^2}{n - 2} =$$

$$\frac{\Sigma(y_i - (a + bx_i))^2}{n - 2} \tag{2.5}$$

An estimate of the variance of the regression coefficient is given by:

$$s_b{}^2 = \frac{s^2}{\Sigma(x_i - \bar{x})^2} \tag{2.6}$$

and this variance has the t-distribution with n-2 degrees of freedom under normal theory. For 95% confidence limits in the present example, we look up the 0.025 ($\alpha/2$) value of t with 8 degrees of freedom, finding it to be 2.306, and calculate:

Upper 95% confidence limit = $b + t(s_b) = 0.516 + 2.306(3.952/630.5)^{1/2} = 0.698$,

and the analogous lower limit is 0.333. Note that these limits are somewhat wider than the 95% limits obtained by bootstrapping. A small sample of quite variable data is involved. It is always important to look at a plot of the data.

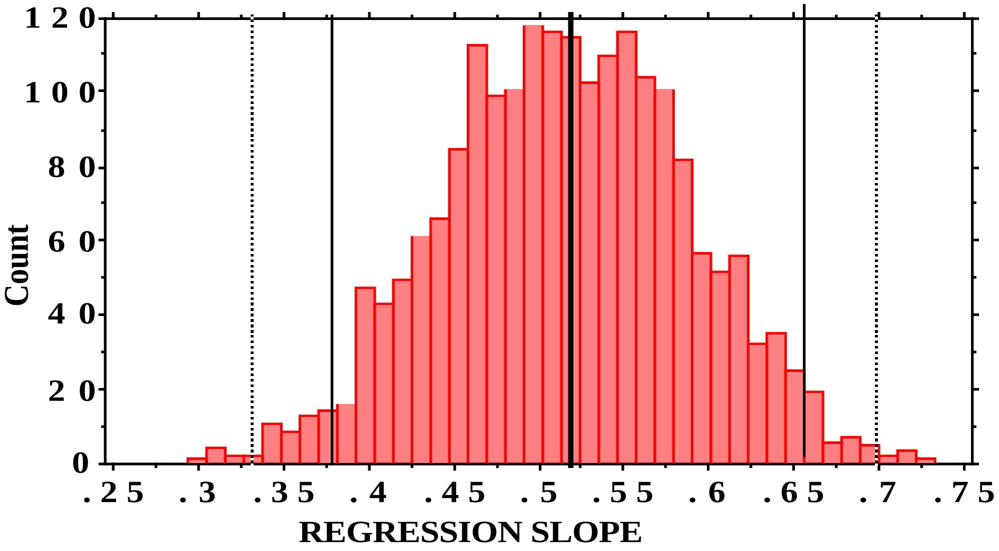

Fig. 2.3. Frequency plot of 2,000 bootstraps created by a BASIC program. The heavy central line shows the position of the regression slope calculated from the original data, while the lighter solid lines show the 95% bootstrap confidence limits. Broken lines show normal 95% confidence limits calculated from the observed data.

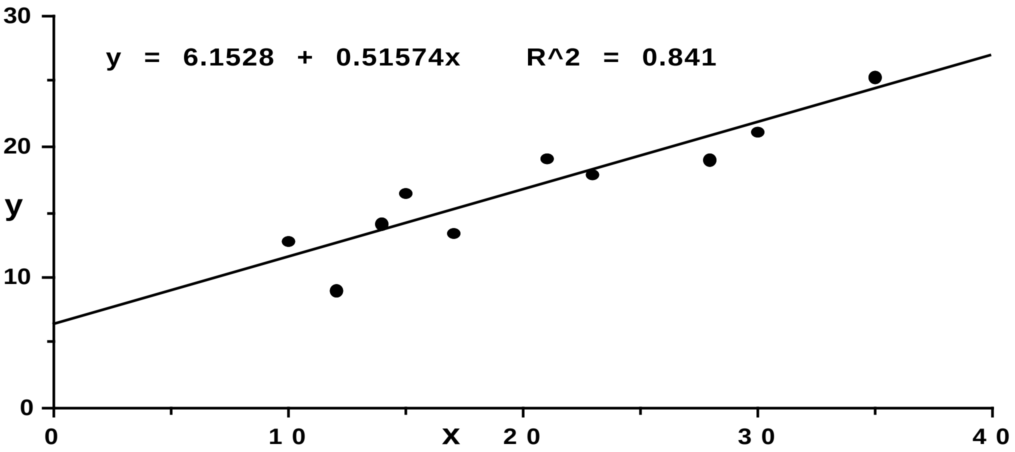

Fig. 2.4. Regression plot of the original data used in Example 2.2.

Most "canned" statistical programs also give the correlation coefficient, which is defined as:

$$r = = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{[\Sigma(x_i - \overline{x})^2\Sigma(y_i - \overline{y})^2]^{1/2}} \qquad (2.7)$$

In the present example, r = 0.917, and is reported by the graphics program that produced Fig. 2.3 as $r^2 = 0.841$. A very serious problem for ecologists is that much of the data they encounter is not normally distributed, and routine use of statistical packages without examining the assumptions or studying the data can lead to important errors in interpreting the data. Bootstrapping provides a way to examine the data without the normal theory assumptions, and thus helps to avoid blunders. The above set of regression data does conform to the normal theory model, so it is worthwhile to look at another example from a different source for contrast. The basis for claiming conformity to normal theory is that the data were constructed using normally distributed errors.

Example 2.3. A regression estimate of survival rate. As a further example, we consider a common use of regression methods. Many investigators are interested in estimating survival rates. Suppose we observe 100 marked animals over 10 years, and tally the number of survivors at the end of each year. If the probability of survival holds constant from year to year and animal to animal, then we can consider that the expected number surviving x years is just $E(n) = Np^x$, where N is the number originally marked and p is the probability of surviving a year. We might then use a model, $y_i = Ns^x$, where y_i is the number observed at the end of the x^{th} year and s is the survival rate. Taking logarithms gives:

$$\log y_i = \log N + x \log s \qquad (2.8)$$

and an easy approach is just to fit a simple linear regression equation, y = a + bx, where b = log s, and use e^b to estimate s. An example of such a data set follows:

Year	Survivors	Log survivors
1	89	4.48864
2	83	4.41884
3	74	4.30407
4	68	4.21951
5	65	4.17439
6	60	4.09435
7	55	4.00733
8	51	3.93183
9	48	3.87120
10	38	3.63759

Plotting log survivors against year gives the following graph:

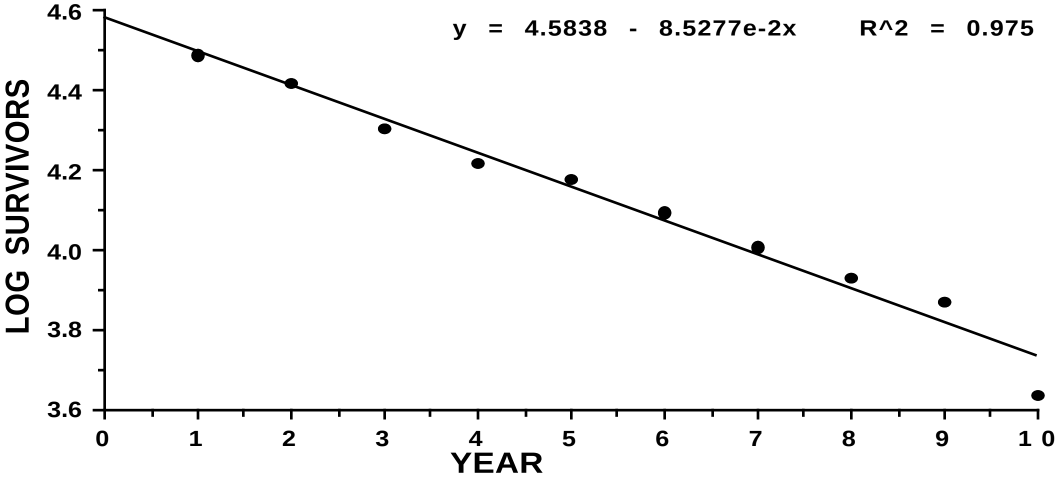

Fig. 2.5. Logarithms of the number of animals surviving at the end of each year regressed against time in years.

Using normal linear regression theory as in the previous example gives a slope of -0.08528 with 95% confidence limits of -0.0742 to -0.0963, and translating these back to a survival rate and confidence limits gives $e^{-0.08528}$ = 0.918 with approximate 95% confidence limits of 0.908 to 0.928. A run of 2,000 bootstraps (parametric regression) gave the following frequency distribution, and 95% confidence limits of -0.09381 and -0.07717, which translate to an annual survival range of 0.910 to 0.926. The bootstrapping was done with a BASIC program, but could have been conducted in EXCEL, just as in the previous example.

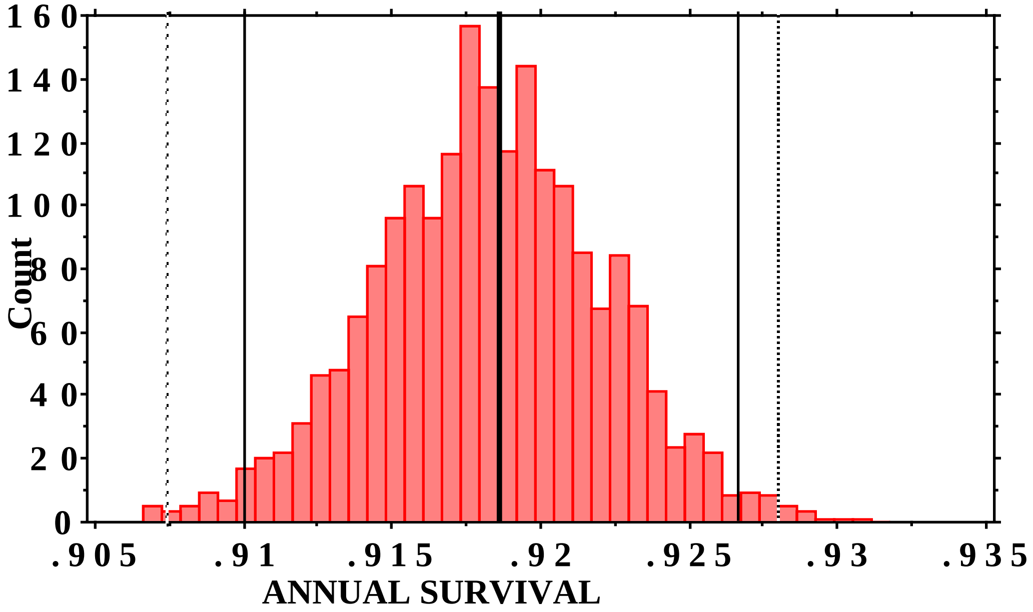

Fig. 2.6. Results of 2,000 parametric regression bootstraps for survival data. Bold central line shows the regression survival estimate (0.928) and solid lines the bootstrap 95% confidence limits. Broken lines are the 95% confidence limits obtained from normal regression theory.

The bootstrap limits appear to be a little "tighter" than the normal theory limits. How can we determine which method is right? One approach is to use "Monte Carlo" methods, which in this case amount to running many stochastic simulations of survival data, and determing which of the two choices for calculating confidence limits gives the best "coverage", i.e., do the calculated confidence limits include the true survival rate in 95% of simulated cases?

Example 2.4. The correlation coefficient. The correlation coefficient (r) calculated as in eq.(2.7) is widely used, along with the assumption that a transformation to:

$$z = 0.5 \log_e \frac{1+r}{1-r} \qquad (2.9)$$

is normally distributed with expected value

$$\mu = 0.5 \log_e \frac{1+\rho}{1-\rho} \text{ and variance } \frac{1}{n-3}.$$

Approximate 95% confidence limits are obtained from $z \pm 2\{\frac{1}{n-3}\}^{1/2}$. Thus in Example 2.2, we had r = 0.917 which is transformed to:

$z = 0.5 \log_e\{\frac{1.917}{.083}\} \pm 2[\frac{1}{7}]^{1/2}$ or $z_1 = 2.326$ and $z_2 = 0.814$. These confidence limits for the transformed variable are usually transformed back by iteritive solutions of eq. (2.9), i.e., we find r_1 and r_2 from:

$$2.326 = 0.5 \log_e \frac{1+r}{1-r} \text{ and } 0.814 = 0.5 \log_e \frac{1+r}{1-r},$$

which gives 95% confidence limits on r as 0.67 to 0.98. If we resort to bootstrapping, then the 2000 bootstraps used to produce Fig. 2.3 (values of r were computed at the same time that values of b were calculated) gave approximate 95% confidence limits of 0.86 to 0.98, essentially the same upper limit, but an appreciably higher lower limit. A graph of the results follows:

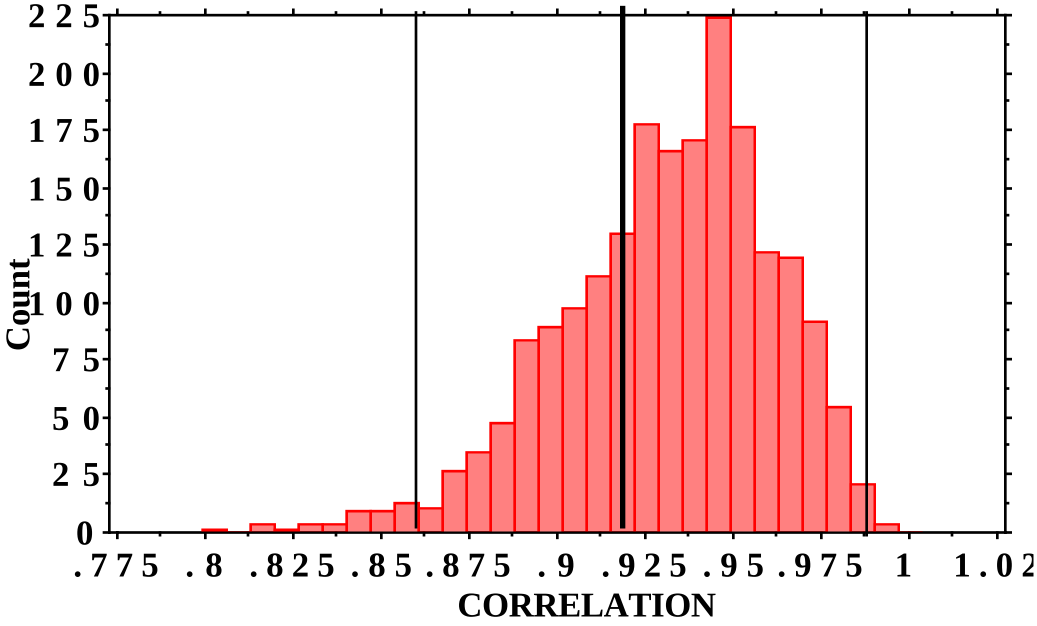

Fig. 2.7. Results of 2,000 bootstraps for the correlation coefficient of the regression data of Example 2.2. Heavy line shows correlation coefficient calculated from the original data, while lighter lines are approximate 95% confidence limits from bootstrapping.

Clearly, the bootstrapped values of the correlation coefficient are quite skewed, but this is the situation with respect to normal regression theory also; the correlation of two jointly normal distributions has a skewed distribution (unless ρ = 0).

2.10 EXERCISES

Where bootstraps of 1,000 or more are involved, students should do the work in individual spreadsheets unless their computer has a large memory. Otherwise you are likely to get an “out of memory” notice when you try to copy from one workbook to another, etc. Make two worksheets for each such exercise, one summarizing results and the second containing the calculations. The practical approach is to save only the first 20 lines or so, when you have finished an exercise. You then can likely consolidate all results in one workbook to hand in. It is important to have your exercises in a workbook, as that makes it possible to try to find out where you went wrong if necessary. **IF YOU WANT TO LEARN TO BOOTSTRAP IN EXCEL, IT IS ESSENTIAL TO DO THE EXERCISES!** The exercises are more or less interlocking so you will need to do most of them. If you do, you should have a pretty fair notion of how to bootstrap. The bootstrapping technique will be used for examples and exercises in the rest of the book, so you need to know how to do it. If you know a programming language, you can certainly do the exercises that way, and provide summary tables and graphs to turn in. An Appendix provides bootstrapping programs as EXCEL “macros” and you can check results with those programs, but you should do the exercises as outlined in Chapters 2 and 3, and then check them with the programs if you want to do so. If you have not used the “graph wizard” function before, you may have trouble getting appropriate x-values on the graph. The trick is to first make an “xy (scatterplot)” graph, finish it and then open the CHART menu and select the bar chart. This changes the xy plot to a bar chart with the appropriate x-axis labels.

2.10.1 Set up an EXCEL worksheet to carry out bootstrap calculations on the data of Fig. 2.1, following the approach outlined in Table 2.1. Use 200 bootstraps. Set up the worksheet to use manual calculation as indicated in Example 2.1 and make 30 runs, recording the mean of the 200 bootstrap means in a separate column (you need to either type in the observed values as you repeatedly run the bootstrapping or use the “PASTE SPECIAL” command). Also calculate the variance of each group of 10 bootstrap samples, listing it at the bottom of the set along with the sum and mean of each set of 10. Record your results on a spreadsheet and save it for the next exercise. Calculate s.e.(boot) of eq. (2.1).

The explanation of using HLOOKUP in EXCEL manuals may not be very helpful. The sample below might help. This is part of a worksheet set up as indicated above and the HLOOKUP string displayed in the header of the worksheet is as follows, referring to the bootstrap sample in the box in the body of the table. It commands EXCEL to lookup the random number in cell D3 (which is 3) in the table of the first two rows (designated in the command as the array $D1:$M$2 and shown in boldface type below) and find the corresponding item in the second row of the array table (which is 203).

=HLOOKUP(D3,$D$1:$M$2,2,FALSE)

A	B	C	D	E	E	G	H	I	J	K	L	M			
EXERCISE			**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**			
2.10.1	ORIGINAL DATA		**13**	**106**	**203**	**131**	**160**	**8**	**67**	**61**	**11**	**301**			
		1	3	9	8	3	7	4	9	1	1	5	10	5	5
		2	2	4	5	2	6	3	9	1	8	7	6	6	9
		3	7	4	1	6	3	2	3	1	3	6	6	1	1
		4	5	8	3	5	9	8	10	7	4	7	10	10	8
	RANDOM NOS.	5	5	2	3	10	10	4	3	8	9	4	5	9	5
		6	8	3	9	4	5	10	4	3	10	6	9	3	8
		7	8	2	8	8	7	9	1	10	6	10	9	8	3
		8	2	5	1	10	5	5	9	2	1	4	7	1	3
		9	7	8	10	3	6	5	2	3	5	10	4	8	5
		10	6	9	2	1	5	6	6	7	6	4	1	8	8
		1	203	11	61	203	67	131	11	13	13	160	301	160	160
		2	106	131	160	106	8	203	11	13	61	67	8	8	11
		3	67	131	13	8	203	106	203	13	203	8	8	13	13
	BOOTSTRAP	4	160	61	203	160	11	61	301	67	131	67	301	301	61
	SAMPLES	5	160	106	203	301	301	131	203	61	11	131	160	11	160
		6	61	203	11	131	160	301	131	203	301	8	11	203	61
		7	61	106	61	61	67	11	13	301	8	301	11	61	203
		8	106	160	13	301	160	160	11	106	13	131	67	13	203
		9	67	61	301	203	8	160	106	203	160	301	131	61	160
		10	8	11	106	13	160	8	8	67	8	131	13	61	61
	BOOTSTRAP		1	2	3	4	5	6	7	8	9	10	11	12	13
		SUM	999	981	1132	1487	1145	1272	998	1047	909	1305	1011	892	1093
		MEAN	99.9	98.1	113	149	115	127	99.8	105	90.9	131	101	89.2	109

The second item in the column of bootstrap samples (just below the item in a box) has the following command:
=HLOOKUP(D4,$D$1:$M$2,2,FALSE)
which instructs EXCEL to find the random number in the position D4 (remember that the first row of the table above, with letters A, B, C etc. is NOT part of a worksheet but merely gives locations in that worksheet). This random number is 2 and thus EXCEL picks out the second item in the array which is 106. The third command is as follows, and EXCEL uses the random number in D5 to pick out the 7th item in the array which is 67.
=HLOOKUP(D5,$D$1:$M$2,2,FALSE)
You may need to exert considerable patience and some trial and error efforts to get EXCEL to do the job if you have not worked much with it before, but once you have the hang of it, things should go along o.k.

2.10.2 Copy the original data and the run of 30 means obtained above to another spreadsheet, and compute means and variances for the two sets of data (the original data, 10 observations and the 30 means) using the built-in functions, i.e., AVERAGE() and VAR(). Make histograms of these means and variances from a run of the spreadsheet made in Exercise 2.10.1 (that is, make histograms of the 200 bootstrap means and variances on that sheet). Show how se_{boot} of eq.(2.1) compares with the variance of the original 10 observations and the variance of the 30 means. Is there any advantage to using bootstrapping in this example?

2.10.3 Repeat the die-tossing example of Section 2.3 using EXCEL. Calculate the expected values. Explain the difference between a p.d.f. and an empirical probability distribution. Which is which in this example? What is the difference between a cumulative distribution function and a distribution function? State the distribution function for this example (as an equation).

2.10.4 Carry out the parametric bootstrap calculations for Example 2.3 using an EXCEL spreadsheet (the approach is given in Example 2.2). Use 1,000 bootstraps, and use VLOOKUP(). It is easier to use for larger numbers of bootstraps. Calculate confidence limits and prepare a graph of the frequency distribution to compare with Fig. 2.7. Use HISTOGRAM to obtain the frequency distribution. Be sure to "freeze" the appropriate cell references using RC so that the x-values remain the same in calculating the bootstraps. You can obtain confidence limits simply by ordering the slopes using the SORT function.

2.10.5 Bootstrap the data of Example 2.2, using 1,000 bootstraps and computing the correlation coefficients rather than the slopes. You can start by copying the bootstrap calculation of Exercise 2.10.4 to a new spreadsheet and inserting the x and y values in this sheet. One can often convert a bootstrap operation to a new data set this way, so it is wise to keep examples on hand. Make a frequency distribution of z (eq.(2.9). Does this look like a normal distribution as assumed in calculating confidence limits under the usual theory?

2.10.6 How would you obtain bootstrap confidence limits on α in Example 2.3? Calculate the 95% bootstrap confidence limits using 1000 bootstraps. Run the EXCEL regression on the data and compare the confidence limits on α with those you obtained from bootstrapping.

2.10. 7 The regression bootstrap of Example 2.3 used parametric bootstrapping in which deviations from a model fitted to the original data are bootstrapped. The first example of bootstrapping given (Example 2.1) might thus be called "nonparametric" bootstrapping. Try this approach on the data of Example 2.3. Remember that you need to bootstrap pairs of observations. This may require setting up the slope calculations in blocks of 10, but careful use of the $function will facilitate copying down in blocks of 10 without too much trouble. Use B = 200, and make a frequency plot of the calculated slopes and compare it with the frequency diagram of Exercise 2.10.4. . This should illustrate why parametric bootstrapping is preferred for small samples in regression studies. This exercise can be time-consuming and illustrates why a programming approach is needed. Try using the program in the Appendix and run 2,000 bootstraps with it.

2.10.8 Referring to the data of Example 2.3, calculate bootstrap 95% confidence limits on the variance about regression as shown in eq. (2.5). Compare your results with the value you get from a regression calculation on the original data of Example 2.3. Report your results on a worksheet (along the lines of those used thus far in the exercises above). You can use the results of Exercise 2.10.6 as a starting point adding on columns containing the sum-of-squares calculation and adding these up to get a variance estimate.